ヒューマンインタフェース 91－2
音 楽 情 報 科 学 37－2
（2000．10．12）

リアルタイム音楽情景記述システム:
全体構想と音高推定手法の拡張

後藤 真孝

電子技術総合研究所
goto@etl.go.jp

あらまし 本稿では，計算機による音楽理解の実現を目指すリアルタイム音楽情景記述システムの全体構想を述べる．これは，人間は楽譜や分離信号を得て音楽を理解していないという立場から，何ができれば音楽を理解したといえるのかを問い直し，「しろうと」が容易にわかるメロディーや楽曲構造のような記述を，音楽CDからリアルタイムに得るシステムを構築する研究アプローチである．そして，既に提案したメロディーとベースの音高推定手法を拡張するアイディアとして，音モデルの多重化，音モデルの推定，音高や音モデルに関する事前分布の導入の三つを提案する．これらをEMアルゴリズムによる最大事後確率推定によって実現し，CDによる実世界の音響信号を用いて実験した結果，音高推定の性能向上が確認された．

A Real-time Music Scene Description System:
System Overview and Extension of F0 Estimation Method

Masataka Goto

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 Japan

Abstract This paper describes an overview of a real-time music scene description system for the computational modeling of music understanding. From a viewpoint that a listener understands music without obtaining musical scores and segregated signals, we consider what is to be achieved for understanding music and build a real-time system that obtains, from music CDs, descriptions intuitive to untrained listeners, such as the melody and the structure of a musical piece. This paper also proposes three ideas (introducing multiple tone models, estimating tone models, and introducing a prior distribution) for extending our method of estimating the F0 of melody and bass lines. Experimental results with the method based on MAP estimation using the EM algorithm show the performance improvement.

1 はじめに

本研究の最終的な目標は，音楽音響信号を人間のように理解できる計算機システムを工学的に実現することである．人間は音楽を聞いて，メロディーを口ずさんだり，曲のビート(拍)に合わせて手拍子を打ったり，同じフレーズが繰り返されていることに気付いたりできる．しかし，人間のこうした音楽理解能力の仕組みは解明されておらず，計算機上で実現するための手法も確立していない．音楽理解能力を備えた計算機システムの構築は，メディア検索や音楽編集等の様々な用途で潜在的に望まれているにも関わらず，困難な課題であるためにまだ実現できていないのが現状である．本研究は，その現状を打破することを狙って，コンパクトディスク(CD)等に収録されている実世界の複雑な音響信号を入力として，リアルタイムに音楽理解が可能なシステムを実現することを目的とする．

音楽音響信号理解へ向けた従来の典型的なアプローチは，自動採譜と音源分離である☆1．しかし，具体的な問題点は2章で指摘するが，音楽理解の実現という観点からは，楽譜や分離信号を得るという問題設定は適切でない．そこで本研究では，「何ができれば音楽を理解したといえるのか」を問い直し，「人間は楽譜や分離信号に基づいて音楽を理解していない」という問題意識の下に，音楽的に訓練されていない「しろうと」の音楽理解の実現を目指す．具体的には，3章で研究の全体構想を述べるように，入力された音楽音響信号に対応した記述を理解結果として出力する**リアルタイム音楽情景記述システム**を構築していく．そして，人間には容易にわかるけれど従来は推定

☆1 個々の文献の紹介は文献1),2)を参照されたい．

困難だと考えられていた，メロディーやベース，フレーズの繰り返し，楽曲構造等のような音楽的要素の記述を，実世界の複雑な混合音から得ることに取り組む．

本稿では，このような全体構想を提案するだけでなく，4章において，その部分問題であるメロディーとベースの**音高**(本稿では**基本周波数**の意味で用いる)を推定するための手法についても述べる．文献3)~5)で提案した手法に対して，各音高に複数の音モデル(高調波構造の確率分布)を用意し，従来固定されていた音モデル自体の推定を可能にした上で，音高や音モデルに関する事前分布を考慮した推定(EMアルゴリズムを用いた最大事後確率推定)ができるように拡張する．そして，5章では，実装した音高推定手法による実験結果を示す．最後に，6章でまとめを述べる．

2 自動採譜と音源分離の問題点

音楽音響信号理解の研究分野で古くから主流であったテーマは，モノラルの音響信号からそれに対応する楽譜を作成する自動採譜である．1970～1980年代には単旋律が主な対象であったが，1980～1990年代には混合音を対象とした研究が増えてきた．そこで必要だと考えられているのが，個々の楽器音の信号を分離抽出する音源分離である．しかし，モノラルの混合音から個々の楽器音を分離する問題は，知覚的な制約が明らかでない不良設定な逆問題で解決が難しく，現状では，高々三つの楽器音が同時に鳴る混合音を不完全に分離できるに止まっている．なお，音源数より多い個数のマイクの出力を信号処理して物理的に音を分離する技術(マイクロホンアレイやICA等)もあるが，原理的にモノラル信号には適用できず，モノラル信号中の音を人間が聞き分けられることからもわかるように，本質的な音楽理解へ向けたアプローチではない．

これらの研究アプローチが抱える問題点は，自動採譜と音源分離が，音楽理解において必要条件でも十分条件でもないということである．

(1) **自動採譜：人間は楽譜化(楽譜上の個々の音符へシンボル化)せずに音楽を理解している．**
楽譜化は，音楽的に訓練された「くろうと」だけが持っているような，獲得困難な高度な技能である．そのような「くろうと」は楽譜に近い分析的な聞き方をしたとしても，訓練されていない「しろうと」は，音楽を聞いても楽譜を思い浮べたりはせず，個々の音符に還元しないで，メロディー等の高次の音楽的要素を把握して音楽を理解している(例えば，和音は音符に還元されず，全体の響きで捉えられている)．また，そうした高次の音楽的要素は，楽譜が得られても明らかでない．

(2) **音源分離：各音源の音響信号を分離抽出するという問題設定が不適切である．**
音に限らず，一般に分離は理解の必要条件ではない．複数の要素が混在した入力が与えられたときに，各要素に固有の特徴を検出すれば，分離しなくても個々の存在が理解できるからである．実際，聴覚心理学の観点からも，人間は音源分離をしていないという指摘がある[6),7)]☆2．つまり，人間が音楽を理解しているからといって音源分離をしているとは限らず，音源分離という困難な逆問題を解ける実体が，人間の脳も含めて，この世に存在していない可能性が高い．

このような問題意識の下に，音楽音響信号理解へ向けた新たな研究アプローチを3章で提案する．

3 リアルタイム音楽情景記述システム構想

本研究では，複数種類の楽器音や歌声を含む音楽音響信号を入力とし，それに対応した記述を理解結果として出力する**リアルタイム音楽情景記述システム**の構築を目指す．**音楽情景記述** (music scene description)☆3とは，音楽演奏中の刻一刻と変化する情景を分析・理解した結果を記述する処理過程である．ここで重要となるのは，何ができれば音楽を理解したといえるのか，つまり，音楽からどのような記述を得るべきなのかという問題設定である(この問いを吟味せずに答えを「楽譜」としたアプローチが自動採譜といえる)．記述には様々な抽象度のものが考えられるが，本研究では，以下の三つの要件を満たすような適度な抽象度を持つ記述を対象とする．

- 直感的：音楽的に訓練されていない「しろうと」が容易に得られる．
- 基礎的：音楽的に訓練された「くろうと」の分析的で詳細な音楽理解を実現する際に礎となる．
- 有用性：様々な応用を実現する際に役に立つ．

☆2 具体的には，「音を完全に分離することを人間の脳は行っていないという証拠がある．」[6)]，「もし音源分離を，混合されている音源すべてを分離する問題と定義するなら，それはヒトがやっていることとはほど遠い．」[7)] のように指摘されている．

☆3 柏野が文献8)で導入した音楽情景分析(音楽音響信号を対象とした聴覚的情景分析)と考え方は近いが，音源分離をおこなうのではなく，情景の記述を得るという目標をより明示的に表す用語として，敢えて音楽情景記述という造語を考案した．柏野の「情景分析としては，信号よりもむしろ記号表現の抽出を問題として設定する方が自然である」[8)] という問題意識は本研究にも共通だが，記号表現として音符を用いていた点が本研究と大きく異なる．

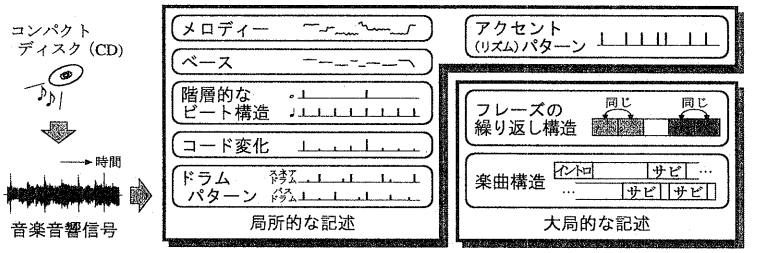

図 1: 8種類の音楽的要素を推定するリアルタイム音楽情景記述システム

3.1 8種類の音楽的要素

以上の要件を考慮して，本稿では，図 1 に示すような8種類の音楽的要素の記述を得る音楽情景記述システムを提案する．8種類の記述は**局所的な記述**と**大局的な記述**に大別され，局所的な記述は基本的に連続値の時間変化として表現される．

(1) **メロディー**

楽曲の中核を担う，他よりも際立って聞こえる単音の系列である．音符列にシンボル化するのではなく，その時間的な変化の軌跡を，音高と振幅の時間変化で表現する．

(2) **ベース**

調性に密接に関連する，アンサンブル中で最も低い単音の系列である．その時間的な変化の軌跡を，音高と振幅の時間変化で表現する．

(3) **階層的なビート構造**

音楽の時間軸方向の局所的な階層構造である．ここでは，人間が音楽中に知覚する基本的な時間単位であるビートのレベル (四分音符レベルと呼ぶ) と小節レベルの二つのレベルを考える☆4．そして，これらを四分音符の時刻 (ビート時刻) の系列と小節の先頭の時刻 (小節時刻) の系列で表現する．

(4) **コード変化**

楽曲の雰囲気に大きい影響を与えるコードが，いつ変化したかという情報である．「しろうと」が得るのが困難なコード名は同定せずに，各時刻においてコードが変化した度合い (曲中のハーモニーがどれくらい変化したと感じるかに対応する度合い) をコード変化度として表現する．

(5) **ドラムパターン**

ドラムスがどのように演奏されたかという情報である．ここでは，ドラムスの中で最も代表的で，リズムに与える影響が大きいバスドラムとスネアドラムを取り上げ，それらが各小節内のどの位置で演奏されたかを示すパターンとして表現する．

(6) **アクセントパターン (リズムパターン)**

アクセント (音量的に大きい点などの知覚的に注意が引かれる点) がどの位置にあったかという情報である．各小節内のどの位置にアクセントが置かれていたかを示すパターンとして表現する．

(7) **フレーズの繰り返し構造**

メロディーの断片 (フレーズ) がどのように繰り返されていたかという情報である．繰り返し演奏される断片のそれぞれに関して，それが楽曲中で再び出現する位置の一覧として表現する．

(8) **楽曲構造**

楽曲全体がどのような構成になっていたかという情報である．楽曲を各構成要素となるブロックごとに分け，それぞれの位置付け (サビであるか，等) を表現する．

以上のシステムを構築していく上で，本研究では，まず最初に実世界の複雑な入力に対して機能するリアルタイムシステムを実現し，次にその能力を向上・拡張していく．これはスケールアップ問題[10] (単純な実験室環境を対象としたシステムは実環境へ拡張困難であるという問題) を考慮したアプローチである．具体的には，我々が普段聞くような現実的な複雑さを持つ，ポピュラー音楽の CD から得たモノラルの音響信号を入力対象とする．CD には通常ステレオ信号が記録されており，仮にステレオ信号を入力とすれば，方向情報に依存した処理が可能で，一般に問題は容易になる．しかし，ステレオ信号を前提としたシステムはモノラル信号に適用できない (逆は適用可能である)．また，人間はモノラル信号からでも上記の音楽的要素を得られる．そこで，本研究では敢えてステレオ信号の左右を混合したモノラル信号を入力とする．

3.2 実現へ向けて

図 1 の全体構想を実現するには，個々の音楽的要素の記述を得る手法の考案・実装と，それらを統合化し，

☆4 本研究は楽譜表現を前提にしていないが，説明の便宜上，楽譜の用語を文献9) のように用いる．例えば，四分音符レベルは人間が音楽中に感じる基本的な時間単位を示すが，これは通常楽譜の四分音符に対応している．

リアルタイムに動作するシステムとして実現するフレームワークの実装をおこなう必要がある．この中で，まずメロディーとベースの推定に関しては，文献3)～5)で提案した音高推定手法 *PreFEst* (Predominant-F0 Estimation Method) を拡張して実現する．具体的な拡張については，4章で議論する．次に，階層的なビート構造，コード変化，ドラムパターンに関しては，文献11)～14)の手法を改良して実現する．他の音楽的要素に関しては，上記の音楽的要素を手がかりとして活用しながら実現するが，具体的な手法は検討段階であり，稿を改めて報告する予定である．

個々の推定手法を統合化したシステムを一つのプロセスとして実装するのは，計算量の多さからも現実的でなく困難である．そこで，個々の機能を別々の計算機上で実行し，通信により情報共有しながら処理を進める．RMCP (Remote Music Control Protocol)[15]は，こうした情報共有と負荷分散に適しており，これを音響信号の伝送用に拡張したRACP (Remote Audio Control Protocol)に基づいて，LAN上に分散した異なるプロセスとして実装する．

4 メロディーとベースの音高推定手法の拡張

PreFEstは，帯域制限された混合音中で最も優勢な音高を推定する手法である．メロディーは中高域において最も優勢な高調波構造を持ち，ベースは低域において最も優勢な高調波構造を持つことが多いため，それぞれに適した帯域に制限して適用すれば，メロディーとベースの音高を推定できる．本手法が対象とするCD等による実世界の音響信号は，事前に音源数を仮定することが不可能な混合音であり，そこでは，周波数成分が頻繁に重複する上に，基本周波数成分が存在しないような音も存在する．しかし，従来の音高推定手法の多くは，少数の音源数を仮定し，周波数成分を局所的に追跡したり，基本周波数成分の存在に依存したりしていたために，そのような実世界の混合音には適用できなかった．それに対して本手法は，音源数を仮定せず，周波数成分の局所的な追跡もおこなわず，基本周波数成分の存在を前提としないという特長を持つ．

そのような特長を実現するために，入力の混合音には，あらゆる基本周波数(本章では以下，音高ではなく，より正確なこの用語を用いる)の音が様々な音量で同時に含まれていると考える．ここでは，統計的手法を利用するために，帯域制限された周波数成分を確率密度関数(観測した分布)で表現し，各音の高調波構造に対応する確率分布を**音モデル**として導入する．そして，周波数成分の確率密度関数が，対象とするあらゆる基本周波数の音モデルの**混合分布モデル**(重み付き和のモデル)から生成されたと考える．この混合分布中の各音モデルの重みは，各高調波構造が相対的にどれくらい優勢かを表すことから，**基本周波数の確率密度関数**と呼ぶ(混合分布中において音モデルが優勢になればなるほど，そのモデルの基本周波数の確率が高くなる)．文献3)～5)で示したように，この重みの値(すなわち基本周波数の確率密度関数)は，EM (Expectation-Maximization)アルゴリズム[16]を用いることで推定できる．そして，求めたい最も優勢な高調波構造の周波数は，単純には，この基本周波数の確率密度関数を最大にする周波数として得られる．しかし，同時に鳴っている音の基本周波数に対応する複数のピークが拮抗すると，それらのピークが次々と選ばれて安定しないため，ピークの時間的な連続性を考慮する必要がある．そのために，マルチエージェントモデルを導入し，複数のエージェントがそれぞれ異なるピークを追跡することで，安定な音高推定結果を得ることができる．

本章では，このような従来のPreFEstを，以下の三点において拡張する方法を提案する．

[拡張1] 音モデルを多重化

従来は同一基本周波数には一つの音モデルしか用意していなかったが，実際には，ある基本周波数に，異なる高調波構造を持つ音が入れ替わり立ち替わり現れることがある．そこで，同一基本周波数に対して複数の音モデルを用意し，それらの混合分布でモデル化する．

[拡張2] 音モデルのパラメータを推定

従来の音モデルでは，各高調波成分の大きさの比率を固定していた(ある理想的な音モデルを仮定していた)．これは実世界の混合音中の高調波構造とは必ずしも一致しておらず，精度向上のためには洗練させる余地が残されていた．そこで，音モデルの高調波成分の大きさの比率もモデルパラメータに加え，各時刻においてEMアルゴリズムで推定する．

[拡張3] モデルパラメータに関する事前分布を導入

従来は音モデルの重み(基本周波数の確率密度関数)に関する事前知識は仮定していなかった．しかし本手法を様々な用途に適用していく上で，たとえ事前に基本周波数がどの周波数の近傍にあるかを与えてでも，より誤検出の少ない基本周波数を求めたいというような応用が考えられる．例えば，演奏分析やビブラート分析等の目的では，楽曲をヘッドホン聴取しながらの歌唱や楽器演奏

によって，各時刻におけるおおよその基本周波数を事前知識として用意しておき，実際の楽曲中のより正確な基本周波数を得ることが求められている．そこで，従来のモデルパラメータの最尤推定の枠組みを拡張し，モデルパラメータに関する事前分布に基づいて最大事後確率推定(MAP推定: Maximum *A Posteriori* Probability Estimation)をおこなう．その際，**[拡張2]**でモデルパラメータに加えた，音モデルの高調波成分の大きさの比率に関する事前分布も導入する．

4.1 拡張方法

帯域制限された周波数成分の確率密度関数 $p_{\Psi}^{(t)}(x)$ から，基本周波数の確率密度関数 $p_{F0}^{(t)}(F)$ を求める過程における，具体的な拡張方法を述べる．本稿で述べる拡張はすべてこの過程にのみ関連するため，紙面の制約から，入力音響信号から $p_{\Psi}^{(t)}(x)$ を得る処理と，$p_{F0}^{(t)}(F)$ からマルチエージェントモデルに基づいて最終出力を決定する処理の詳細に関しては，文献3)~5)に譲って省略する．ここで，上記の t はフレームシフト(10 msec)を時間単位とする時刻，x と F はcentの単位☆5で表された対数スケールの周波数である．

4.1.1 拡張した音モデルの混合分布モデル

まず，**[拡張1]**と**[拡張2]**を実現するために，同一基本周波数に対して M_i 種類の音モデルがあるものとし(i はメロディー用 ($i=\mathrm{m}$) かベース用 ($i=\mathrm{b}$) かを示す)，基本周波数が F の m 番目の音モデルの確率密度関数 $p(x|F,m,\mu^{(t)}(F,m))$ にモデルパラメータ $\mu^{(t)}(F,m)$ を導入する．これらは，次式のように表されるものとする(図2)．

$$p(x|F,m,\mu^{(t)}(F,m)) = \sum_{h=1}^{\mathrm{H}_i} p(x,h|F,m,\mu^{(t)}(F,m)) \quad (1)$$

$$p(x,h|F,m,\mu^{(t)}(F,m)) = c^{(t)}(h|F,m)\, G(x;F+1200\log_2 h,\mathrm{W}_i) \quad (2)$$

$$\mu^{(t)}(F,m) = \{c^{(t)}(h|F,m) \mid h=1,...,\mathrm{H}_i\} \quad (3)$$

$$G(x;x_0,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-x_0)^2}{2\sigma^2}} \quad (4)$$

これは，基本周波数が F のときに，その高調波成分がどの周波数にどれくらい現れるかをモデル化したものである．H_i は基本周波数成分も含めた高調波成分の数，W_i^2 はガウス分布 $G(x;x_0,\sigma)$ の分散を表す．$c^{(t)}(h|F,m)$ は，第 h 次高調波成分の大きさを表し，次式を満たすものとする．

$$\sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h|F,m) = 1 \quad (5)$$

そして，周波数成分の確率密度関数 $p_{\Psi}^{(t)}(x)$ が，次式で定義されるような，$p(x|F,m,\mu^{(t)}(F,m))$ の混合分布モデル $p(x|\theta^{(t)})$ から生成されたと考える．

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m)\, p(x|F,m,\mu^{(t)}(F,m))\, dF \quad (6)$$

$$\theta^{(t)} = \{w^{(t)},\mu^{(t)}\} \quad (7)$$

$$w^{(t)} = \{w^{(t)}(F,m) \mid \mathrm{Fl}_i \le F \le \mathrm{Fh}_i, m=1,...,\mathrm{M}_i\} \quad (8)$$

$$\mu^{(t)} = \{\mu^{(t)}(F,m) \mid \mathrm{Fl}_i \le F \le \mathrm{Fh}_i, m=1,...,\mathrm{M}_i\} \quad (9)$$

ここで，Fh_i と Fl_i は，許容される基本周波数の上限と下限であり，$w^{(t)}(F,m)$ は，次式を満たすような，音モデルの重みである．

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m)\, dF = 1 \quad (10)$$

実世界の混合音に対して事前に音源数を仮定することは不可能なため，式(6)のように，あらゆる基本周波数の可能性を同時に考慮してモデル化することが重要となる．最終的に，モデル $p(x|\theta^{(t)})$ から，観測した確率密度関数 $p_{\Psi}^{(t)}(x)$ が生成されたかのようにモデルパラメータ $\theta^{(t)}$ を推定できれば，その重み $w^{(t)}(F,m)$ は各高調波構造が相対的にどれくらい優勢かを表すため，次式のように基本周波数の確率密度関数 $p_{F0}^{(t)}(F)$ と解釈することができる．

$$p_{F0}^{(t)}(F) = \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F,m) \quad (\mathrm{Fl}_i \le F \le \mathrm{Fh}_i) \quad (11)$$

☆5 Hzで表された周波数 f_{Hz} は，$f_{\mathrm{cent}} = 1200\log_2 \frac{f_{\mathrm{Hz}}}{440\times 2^{\frac{3}{12}-5}}$ により centで表された周波数 f_{cent} に変換されるものとする．

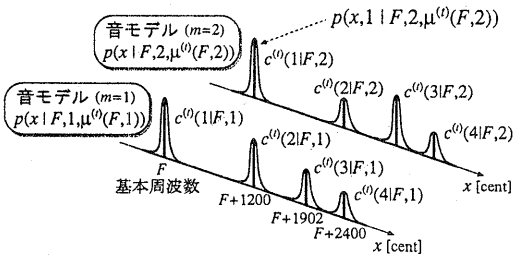

図 2: 音モデルのパラメータの推定

4.1.2 事前分布の導入

次に，**[拡張3]**を実現するために，$\theta^{(t)}$ の事前分布 $p_{0i}(\theta^{(t)})$ を，式(12)のように式(13)と式(14)の積で与える．この $p_{0i}(w^{(t)})$ と $p_{0i}(\mu^{(t)})$ は，最も起りやすいパラメータを $w_{0i}^{(t)}(F,m)$ と $\mu_{0i}^{(t)}(F,m)$ ($c_{0i}^{(t)}(h|F,m)$) としたときに，そこで最大値を取るような単峰性の事前分布である．ただし，Z_w, Z_μ は正

規化係数，$\beta_{wi}^{(t)}$, $\beta_{\mu i}^{(t)}(F,m)$ は，最大値をどれくらい重視した事前分布とするかを決めるパラメータで，0のときに無情報事前分布（一様分布）となる．

$$p_{0i}(\theta^{(t)}) = p_{0i}(w^{(t)})\ p_{0i}(\mu^{(t)}) \tag{12}$$

$$p_{0i}(w^{(t)}) = \frac{1}{Z_w} e^{-\beta_{wi}^{(t)}\ D_w(w_{0i}^{(t)};w^{(t)})} \tag{13}$$

$$p_{0i}(\mu^{(t)}) = \frac{1}{Z_\mu} e^{-\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \beta_{\mu i}^{(t)}(F,m)\ D_\mu(\mu_{0i}^{(t)}(F,m);\mu^{(t)}(F,m))\ dF} \tag{14}$$

$D_w(w_{0i}^{(t)};w^{(t)})$, $D_\mu(\mu_{0i}^{(t)}(F,m);\mu^{(t)}(F,m))$ は，次のようなK-L情報量 (Kullback-Leibler's information) である．

$$D_w(w_{0i}^{(t)};w^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F,m)\ log\frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)}\ dF \tag{15}$$

$$D_\mu(\mu_{0i}^{(t)}(F,m);\mu^{(t)}(F,m)) = \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m)\ log\frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} \tag{16}$$

4.1.3 EMアルゴリズムによる最大事後確率推定

以上から，確率密度関数 $p_\Psi^{(t)}(x)$ を観測したときに，そのモデル $p(x|\theta^{(t)})$ のパラメータ $\theta^{(t)}$ を，事前分布 $p_{0i}(\theta^{(t)})$ に基づいて推定する問題を解けばよいことがわかる．この事前分布に基づく $\theta^{(t)}$ の最大事後確率推定量(MAP推定量)は，次式を最大化することで得られる．

$$\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ (\log p(x|\theta^{(t)}) + \log p_{0i}(\theta^{(t)}))\ dx \tag{17}$$

この最大化問題は解析的に解くことが困難なため，EMアルゴリズム[16]を用いて $\theta^{(t)}$ を推定する．EMアルゴリズムは，不完全な観測データから最尤推定をおこなうために用いられることが多いが，文献16)にも述べられているように，最大事後確率推定の場合にも適用できる．最尤推定では，平均対数尤度の条件付き期待値を求めるEステップ (expectation step) とその最大化をおこなうMステップ (maximization step) を交互に繰返すが，最大事後確率推定の場合には，条件付き期待値に事前分布の対数を加えたものの最大化を繰り返す．ここでは各繰返しにおいて，古いパラメータ推定値 $\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$ を更新して新しいパラメータ推定値 $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$ を求めていく．

周波数 x において観測した各周波数成分が，どの基本周波数のどの音モデルのどの倍音から生成されたのかを表す隠れ変数 F, m, h を導入して，EMアルゴリズムを以下のように定式化することができる．

1. （Eステップ）

最尤推定の場合には，平均対数尤度の条件付き期待値 $Q(\theta^{(t)}|\theta'^{(t)})$ を求めるが，最大事後確率推定の場合には，それに $\log p_{0i}(\theta^{(t)})$ を加えた $Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)})$ を求める．

$$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}) = Q(\theta^{(t)}|\theta'^{(t)}) + \log p_{0i}(\theta^{(t)}) \tag{18}$$

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ \mathrm{E}_{F,m,h}[\log p(x,F,m,h|\theta^{(t)}) \mid x,\theta'^{(t)}]\ dx \tag{19}$$

ここで，条件付き期待値 $\mathrm{E}_{F,m,h}[a|b]$ は，条件 b により決定される確率分布を持つ隠れ変数 F, m, h に関する，a の期待値を意味する．

2. （Mステップ）

$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)})$ を $\theta^{(t)}$ の関数として最大化して，更新後の新しい推定値 $\overline{\theta^{(t)}}$ を得る．

$$\overline{\theta^{(t)}} = \underset{\theta^{(t)}}{\mathrm{argmax}}\ Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}) \tag{20}$$

Eステップにおいて，式 (19)は

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i}\sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\ p(F,m,h|x,\theta'^{(t)}) \log p(x,F,m,h|\theta^{(t)}) dF dx \tag{21}$$

となる．この式中の完全データの対数尤度は

$$\log p(x,F,m,h|\theta^{(t)}) = \log(w^{(t)}(F,m)\ p(x,h|F,m,\mu^{(t)}(F,m))) \tag{22}$$

で与えられる．また，$\log p_{0i}(\theta^{(t)})$ は，

$$\log p_{0i}(\theta^{(t)}) = -\log Z_w Z_\mu - \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \left(\beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)\ \log\frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m) \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m)\ \log\frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} \right) dF \tag{23}$$

となる．

次に，Mステップに関しては，式 (20)が，式 (5)と式 (10)を条件とする条件付き変分問題となっている．この問題は，Lagrangeの乗数 λ_w, λ_μ を導入し，次のEuler-Lagrangeの微分方程式によって解くことができる．

$$\frac{\partial}{\partial w^{(t)}} \left(\int_{-\infty}^{\infty} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\ p(F,m,h|x,\theta'^{(t)}) (\log w^{(t)}(F,m) + \log p(x,h|F,m,\mu^{(t)}(F,m)))\ dx - \beta_{wi}^{(t)}\ w_{0i}^{(t)}(F,m)\ \log\frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} - \lambda_w(w^{(t)}(F,m) - \tfrac{1}{\mathrm{M}_i(\mathrm{Fh}_i-\mathrm{Fl}_i)}) \right) = 0 \tag{24}$$

$$\frac{\partial}{\partial c^{(t)}} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ p(F,m,h|x,\theta'^{(t)}) (\log w^{(t)}(F,m) + \log c^{(t)}(h|F,m) + \log G(x;F+1200\log_2 h,\mathrm{W}_i))\ dx \right.$$

$$- \beta_{\mu i}^{(t)}(F,m)\, c_{0i}^{(t)}(h|F,m)\, \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} - \lambda_\mu (c^{(t)}(h|F,m) - \tfrac{1}{\mathrm{H}_i}) \Big) = 0 \quad (25)$$

これより，

$$w^{(t)}(F,m) = \frac{1}{\lambda_w} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, p(F,m|x,\theta'^{(t)})\, dx + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m) \right) \quad (26)$$

$$c^{(t)}(h|F,m) = \frac{1}{\lambda_\mu} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m,h|x,\theta'^{(t)})\, dx + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m) \right) \quad (27)$$

が得られる．これらの式において，Lagrangeの乗数は式 (5)，式 (10)から

$$\lambda_w = 1 + \beta_{wi}^{(t)} \quad (28)$$

$$\lambda_\mu = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m|x,\theta'^{(t)})\, dx + \beta_{\mu i}^{(t)}(F,m) \quad (29)$$

と定まり，$p(F,m,h|x,\theta'^{(t)})$, $p(F,m|x,\theta'^{(t)})$ はベイズの定理から，

$$p(F,m,h|x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)\, p(x,h|F,m,\mu'^{(t)}(F,m))}{p(x|\theta'^{(t)})} \quad (30)$$

$$p(F,m|x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)\, p(x|F,m,\mu'^{(t)}(F,m))}{p(x|\theta'^{(t)})} \quad (31)$$

となる．以上から，新しいパラメータ推定値 $\overline{w^{(t)}(F,m)}$ と $\overline{c^{(t)}(h|F,m)}$ を求める式は次のようになる．

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1 + \beta_{wi}^{(t)}} \quad (32)$$

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}\ \overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m)} \quad (33)$$

式中の $\overline{w_{\mathrm{ML}}^{(t)}(F,m)}$ と $\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)}$ は，$\beta_{wi}^{(t)} = 0$, $\beta_{\mu i}^{(t)}(F,m) = 0$ の無情報事前分布のとき，つまり最尤推定の場合の推定値である．

$$\overline{w_{\mathrm{ML}}^{(t)}(F,m)} = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\, p(x|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta} dx \quad (34)$$

$$\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} = \frac{1}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}} \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\, p(x,h|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta} dx \quad (35)$$

これらの反復計算により，事前分布を考慮した基本周波数の確率密度関数 $p_{F0}^{(t)}(F)$ が，式 (11)によって $w^{(t)}(F,m)$ から求まる．さらに，すべての音モデル $p(x|F,m,\mu^{(t)}(F,m))$ の各高調波成分の大きさの比率 $c^{(t)}(h|F,m)$ も求まり，**[拡張 1]**～**[拡張 3]**が実現された．

5 音高推定手法の実装と実験結果

4 章で拡張した手法に基づいて，リアルタイムに動作するメロディーとベースの音高推定システムを分散環境で実装した．音高推定手法の計算はワークステーション (Alpha21264 750 MHz dual CPU, Linux 2.2) 上で実行され，音響信号の入出力や視覚化の処理は別のワークステーション (SGI Octane R12000 300 MHz dual CPU, Irix 6.5) 上で実行される．

表 1の左側に示すポピュラー音楽，ジャズ，クラシックの楽曲10曲からの抜粋を用いて実験をおこなった．入力は市販のCDからサンプリングしたモノラルの音響信号で，それぞれが単音のメロディーと複数種類の楽器音を含んでいる．システムの出力結果の正誤を判定できるように，基準となる正解のメロディーとベースの音高を，人間が手作業で1フレーム (10 msec)ごとに指定するための音高情報エディタを開発した[5]．そして，作成した正解とシステムの出力の周波数差が50 cent以下のときに正しいと判定して，検出率 (正しかったフレームの割合) を求めた．ただし，メロディーやベースが鳴っていないフレームは評価対象外とした．

実験条件を変えながらシステムの検出率を求めた結果を表 1 の右側に示す．条件1は，4 章の拡張を実施する前のPreFEstの性能を示す．パラメータは，$\mathrm{Fh_m}$ = 8400 cent, $\mathrm{Fl_m}$ = 3600 cent, $\mathrm{M_m}$ = 1, $\mathrm{H_m}$ = 16, $\mathrm{W_m}$ = 17 cent, $\mathrm{Fh_b}$ = 4800 cent, $\mathrm{Fl_b}$ = 1000 cent, $\mathrm{M_b}$ = 1, $\mathrm{H_b}$ = 6, $\mathrm{W_b}$ = 17 cent とし，音モデルは $c^{(t)}(h|F,1) = \alpha\, G(h;1,\mathrm{U}_i)$, ($\mathrm{U_m}$ = 5.5, $\mathrm{U_b}$ = 2.7) で固定とした (α は正規化係数)．

それに対して，**[拡張 1]**により音モデルを一つ追加した場合 ($\mathrm{M_m}$ = 2, $\mathrm{M_b}$ = 2) の結果を，条件2に示す．追加した音モデルは，条件1の音モデルの偶数次の高調波成分を $\frac{2}{3}$ 倍に減衰させたものとした．表 1より，曲番号7, 8, 10においてベースの検出率が大きく向上していることがわかる (最大17.6%)．

さらに，音モデルを二つに拡張した状態に加えて，**[拡張 2]**と**[拡張 3]**により，音モデルの事前分布を仮定し，音モデルのパラメータも推定した場合の結果を，条件3に示す．なお，音モデルのパラメータも推定する場合には，モデルの自由度が高いため，このように事前分布を仮定することが不可欠である．音モ

表 1: メロディーとベースの検出率 [%]

曲番号. タイトル (アーティスト名等) [ジャンル]	実験条件					
	1. 拡張前		2. 音モデルを追加		3. 音モデルを推定	
	melody	bass	melody	bass	melody	bass
1. Always *(Bon Jovi)* [ポピュラー]	94.1	85.2	94.0	85.4	94.2	85.4
2. Time Goes By *(Every Little Thing)* [ポピュラー]	91.0	74.3	91.2	73.3	91.6	73.1
3. 星の降る丘 *(Misia)* [ポピュラー]	89.1	76.1	**90.8**	73.0	91.7	72.1
4. My Heart Will Go On *(Celine Dion)* [ポピュラー]	91.3	92.8	91.0	91.2	90.8	91.2
5. Spirit of Love *(Sing Like Talking)* [ポピュラー]	90.9	76.2	90.3	76.9	90.1	76.4
6. Vision of Love *(Mariah Carey)* [ポピュラー]	76.6	88.8	76.2	88.3	76.6	87.3
7. Scarborough Fair *(Herbie Hancock)* [ジャズ]	95.2	55.9	95.4	[**63.1**]	95.3	**65.8**
8. On Green Dolphin Street *(Miles Davis)* [ジャズ]	92.3	51.6	92.8	[**69.2**]	92.9	[**80.8**]
9. Autumn Leaves *(Julian "Cannonball" Adderley)* [ジャズ]	81.9	85.6	82.3	83.8	82.2	82.0
10. Violin Concerto in D, Op. 35 *(Tchaikovsky)* [クラシック]	76.0	66.7	**78.6**	[**78.9**]	78.7	[**84.8**]
平均	87.8	75.3	**88.3**	**78.3**	88.4	**79.9**

太字の検出率は，一つ左の条件より1%以上性能向上が得られたものを示し，
[ ]のついた検出率は，その中でも特に大きな向上が得られたものを示す．

デルの事前分布は，条件2の各音モデルと同一とし，$\beta_{\mu i}^{(t)}(F,m) = \mathrm{B_i}\, e^{-(\frac{F-\mathrm{Fl_i}}{\mathrm{Fh_i}-\mathrm{Fl_i}})^2/0.2}$, ($\mathrm{B_m} = 15$, $\mathrm{B_b} = 10$) とした．その結果，条件2と検出率がほぼ同じ曲が多いものの，曲番号8, 10においてベースの検出率の向上が得られた (最大11.6%)．

今回の拡張では，メロディーの検出率に関しては，大きな性能向上は得られなかった．これは，そもそも条件1でも10曲中6曲の検出率が90%を越えており，ベースよりも性能向上の余地が小さかったことが影響していると考えられる．今後，さらに性能を向上させるためには，音源同定手法を導入して音源の種類等も手がかりに加えていく必要がある．

なお，今回拡張した枠組みは様々な可能性を持っており，上記の結果はあくまで一適用例に過ぎない．例えば，音モデルを多数用意するなどして，今後その枠組みをさらに活用していきたい．また，[**拡張3**]では，基本周波数の確率密度関数に関する事前分布を導入したが，紙面の制約から実験結果まで言及できなかった．これについては稿を改めて報告する予定である．

6 おわりに

本稿では，まず，音楽音響信号理解を目指した研究アプローチとして，リアルタイム音楽情景記述システムの全体構想について述べた．これは，自動採譜や音源分離をせずに，CDによる実世界の音響信号を理解するリアルタイムシステムを構築する新たな試みである．次に，メロディーとベースの基本周波数を推定する手法PreFEstを拡張するための三つのアイディアを提案し，音モデルの多重化や，音モデルの推定，音高や音モデルに関する事前分布を考慮した最大事後確率推定を可能にした．これらの拡張を施した音高推定システムを用いて実験した結果，音モデルの多重化や音モデルの推定が，特にベースの検出率の向上に効果的であることが確認された．

今後は，提案したリアルタイム音楽情景記述システムの実現へ向けて研究を進めていくと共に，PreFEstのさらなる改良をおこなっていく予定である．

謝辞

本研究に対し有益な議論をして頂いた，赤穂 昭太郎 氏，麻生 英樹 氏，速水 悟 氏に感謝する．

参考文献

[1] 片寄晴弘: 自動採譜, コンピュータと音楽の世界, 共立出版, pp. 74–88 (1998).

[2] 柏野邦夫: 重なり合った音を聞き分ける — 音源分離, コンピュータと音楽の世界, 共立出版, pp. 89–99 (1998).

[3] 後藤真孝: 実世界の音楽音響信号を対象としたメロディーとベースの音高推定, 情処研報 音楽情報科学 99-MUS-31-16, pp. 91–98 (1999).

[4] Goto, M.: A Robust Predominant-F0 Estimation Method for Real-time Detection of Melody and Bass Lines in CD Recordings, *Proc. of ICASSP 2000*, pp. II–757–760 (2000).

[5] 後藤真孝: 音楽音響信号を対象としたメロディーとベースの音高推定 (in press), 信学論 (D-II), Vol. J84-D-II, No. 1 (2001).

[6] Bregman, A. S.: 聴覚の情景分析とは (邦訳: 河原 英紀), 日本音響学会誌, Vol. 50, No. 12, pp. 1007–1010 (1994).

[7] 柏野牧夫: 雑音と音声の知覚, 音講論集 秋季 2-2-1 (1998).

[8] 柏野邦夫: 音楽音響信号を対象とする聴覚的情景分析に関する研究, 博士論文, 東京大学 工学部 (1994).

[9] Rosenthal, D.: *Machine Rhythm: Computer Emulation of Human Rhythm Perception*, PhD Thesis, Massachusetts Institute of Technology (1992).

[10] Kitano, H.: Challenges of Massive Parallelism, *Proceedings of IJCAI-93*, pp. 813–834 (1993).

[11] 後藤真孝, 村岡洋一: ビートトラッキングシステムの並列計算機への実装 — AP1000によるリアルタイム音楽情報処理—, 情処学論, Vol. 37, No. 7, pp. 1460–1468 (1996).

[12] 後藤真孝, 村岡洋一: 音響信号を対象としたリアルタイムビートトラッキングシステム— コード変化検出による打楽器音を含まない音楽への対応 —, 信学論 (D-II), Vol. J81-D-II, No. 2, pp. 227–237 (1998).

[13] Goto, M. and Muraoka, Y.: Real-time Beat Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions, *Speech Communication*, Vol. 27, No. 3–4, pp. 311–335 (1999).

[14] 後藤真孝: 音楽音響信号を対象としたリアルタイムビートトラッキングに関する研究, 博士論文, 早稲田大学 理工学部 (1998).

[15] 後藤真孝, 根山亮, 村岡洋一: RMCP: 遠隔音楽制御用プロトコルを中心とした音楽情報処理, 情処学論, Vol. 40, No. 3, pp. 1335–1345 (1999).

[16] Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. B*, Vol. 39, No. 1, pp. 1–38 (1977).